LEARNING IMAGE SUPER RESOLUTION FROM JOINT EXAMPLES

BY

ZHANGYANG WANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Thomas S. Huang

ABSTRACT

Image super-resolution (SR) aims to estimate of a high-resolution (HR) image from low-resolution (LR) input. Image priors are commonly learned to regularize the ill-posed SR problem, either using external LR-HR pairs or internal similar patterns repeating across different scales. We propose joint SR to adaptively combine the advantages of both external and internal SR. We define the two loss functions using sparse coding and epitomic matching, respectively. A corresponding adaptive weight is constructed to balance their effect according to the reconstruction errors. Various image results demonstrate the effectiveness of the proposed method over the existing state-of-the-art methods, which is also verified by our subject evaluation experiment.

To my parents, for their consistent love and support

ACKNOWLEDGMENTS

Thanks to my adviser, Prof. Thomas S. Huang, who guided my graduate research. Thanks to the colleagues and alumni in Image Formulation and Processing (IFP) group, for heated discussions on research almost everyday. They include but are not limited to: Dr. Jianchao Yang, Dr. Zhaowen Wang, Yingzhen Yang, Wei Han, Xianming Liu, Shiyu Chang, Thomas Paine, Pooya Khorrami, and others.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Super-resolution (SR) algorithms aim to construct a high-resolution (HR) image from one or multiple low-resolution (LR) input frames [1]. However, this problem is essentially ill-posed because much information is lost in the HR to LR degradation process. Thus SR has to rely on strong image priors for robust estimation. Such image priors range from the simplest analytical smoothness assumptions, to more sophisticated statistical and structural priors learned from natural images [2], [3], [4], [5].

The most popular single image SR methods rely on example-based learning techniques. Classical example-based methods learn the mapping between LR and HR image patches, from a large and representative external set of image pairs, thus denoted as *external SR*. Meanwhile, images generally possess a great amount of self-similarities; such a self-similarity property motivates a series of *internal SR* methods. With much progress being made, it is recognized that external and internal SR methods each suffer from their certain bottlenecks. However, their complementary properties inspired us to propose the *Joint Super Resolution* (Joint SR), that adaptively utilizes both external and internal examples for the SR task. The contributions of this thesis are multi-fold:

- We propose *joint SR* exploiting both external and internal examples, by defining an adaptive combination of different loss functions.
- We apply *epitomic matching* to enforce self-similarity in SR. It features a robustness to outlier features, as well as its ability to perform efficient non-local searching.
- We carry out a human *subject review* to evaluate SR quality based on visual perception, in addition to conventional quantitative and qualitative comparisons.

The remainder of the thesis is organized as follows. Chapter 2 first studies the motivation and related literature for joint SR. We then develop our joint SR model and discuss each of its components in Chapter 3. Chapter 4 presents implementation details and results of our algorithm, with comparison to a few state-of-the-art methods. Finally, we conclude the thesis in Chapter 5.

CHAPTER 2

WHY JOINT SR: A MOTIVATION STUDY

2.1 External and Internal SR Methods

External SR methods use a universal set of example patches to predict the missing (high-frequency) information for the HR image. In [6], during the training phase, LR-HR patch pairs are collected. Then in the test phase, each input LR patch is found with a nearest neighbor (NN) match in the LR patch pool, and its corresponding HR patch is selected as the output. It is further formulated as a kernel ridge regression (KRR) in [7]. More recently, a popular class of external SR methods are associated with the *sparse coding* technique [8], [9]. The patches of a natural image can be represented as a sparse linear combination of elements in a redundant pre-trained dictionary. The advanced *coupled sparse coding* is further proposed in [4], [9]. External SR methods are known for their capabilities to produce plausible image appearances. However, there is no guarantee that an arbitrary input patch can be well matched or represented by the given external database. Especially when dealing with some unique features of the input image, most external SR methods are prone to producing noise and irregularities [10]. It constitutes the inherent problem of any external SR method with a finite-size training set [11].

Another source of example patches is to search within the input image itself, based on the fact that patches often tend to recur within the image [12], [13], [10], or across different image scales [5]. Although internal examples provide a limited number of references, they are very relevant to the input image. However, this type of approach has a limited performance, especially for irregular patches without any noticeable repeating pattern [14]. Also, due to the unavailability of sufficient patch pairs, the mismatches of internal examples often lead to more severe visual artifacts.

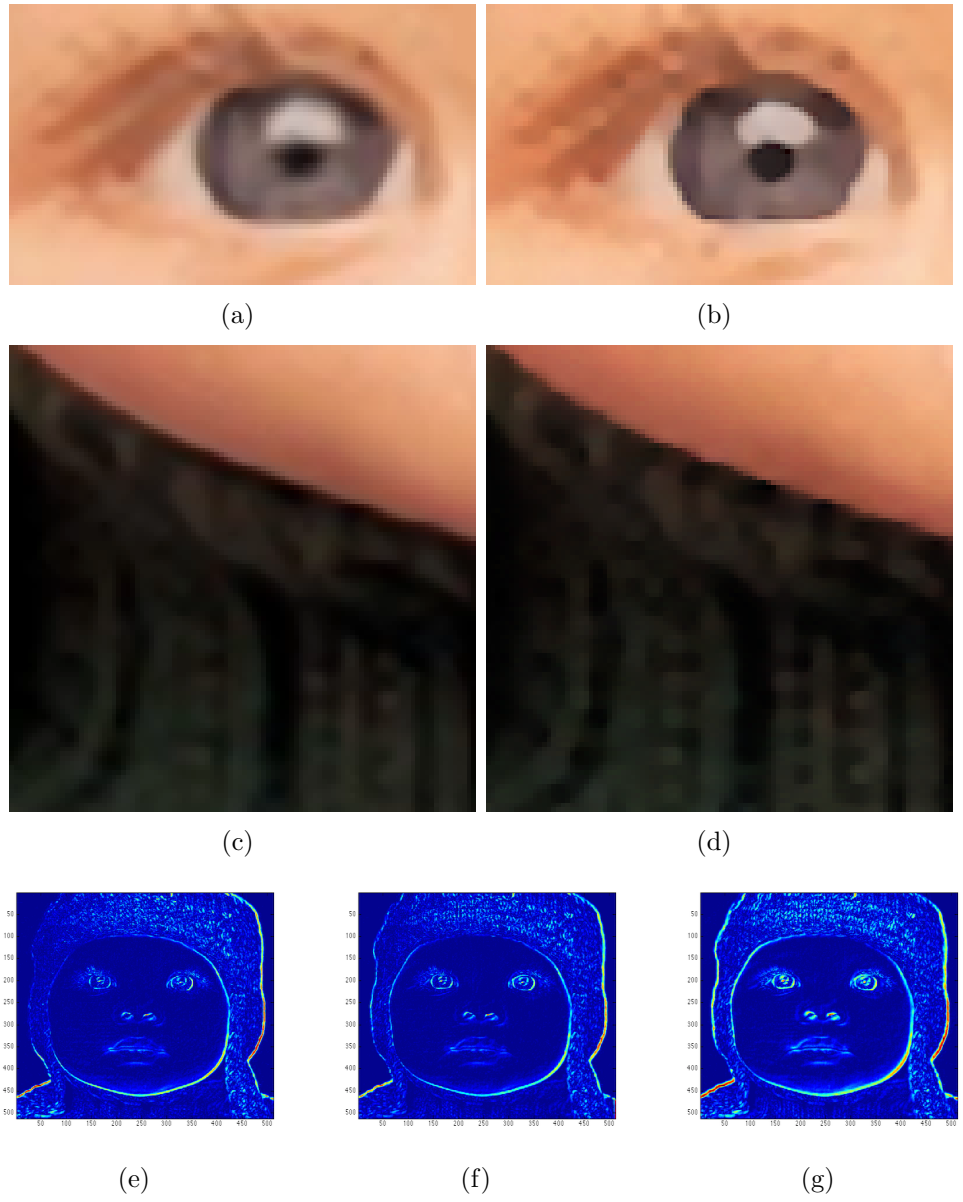


Figure 2.1: (a) Eye region by [4]; (b) eye region by [5]; (c) sweater region by [4], (d) sweater region by [5]. (e) The difference heat map of the SR result by [4] and the ground truth. (f) The difference heat map of the SR result by [5] and the ground truth. (g) The difference heat map of the SR result by bicubic interpolation and the ground truth.

Both external and internal SR methods have different advantages in performing SR. Furthermore, their unique characteristics imply that superior SR results may be expected via a proper integration of the two. See Fig. 2.1 for a specific example: images (a) and (c) are from the 4× SR result of the *Kid* image, by external SR [4], while (b) and (d) are from the 4× result of the same image but with internal SR [5]. Images (a) and (b) are cropped from the same spatial location, and so are images (c) and (d).

First, comparing (a) with (b) shows a winning case of external examples. Although in (b), the black dot and the glare in the eye looks a little sharper, it is easy to notice inconsistent artifacts along the eyelid and structural distortions on the eyeball in (b). That is because the eye region is composed of complex curves and fine structures, where few recurring patterns can be identified in the same image. In contrast, external examples collect references from an external dataset and perform a more natural-looking SR. Thus, internal examples generate sharper SR results in images (d) than (c), since the sweater textures repeat their own patterns frequently, and thus the local neighborhood is rich in internal examples. The two groups of comparisons clearly manifest the different behaviors of external and internal examples for SR. The SR results of the entire *Kid* image are available in Chapter 4.

Figure 2.1 images (e)-(g) display the difference maps between ground truth (original HR *Kid* image), and SR results by bicubic interpolation, external SR and internal SR, in the form of heat maps (using the "jet" colormap), respectively. Although both generate significantly fewer errors than the bicubic interpolation result in (g), the external SR result in (e) tends to produce fewer errors on the face, especially eye regions, whereas the internal SR result (f) shows more favorably in (repeatedly) textured regions.

2.2 Relevant Literature

Based on the observation in Section 2.1, we expect that the external examples contribute to visually pleasant SR results for smooth regions as well as some irregular structures. Meanwhile, internal examples serve as a powerful source to reproduce unique and singular features that rarely appear externally but recur in the same image. Note that similar arguments have been validated statistically in the image denoising literature [15].

There are existing efforts to combine both external and internal examples for image enhancement [16]. Mosseri et al. [17] first proposed that some image patches inherently prefer internal examples for denoising, whereas other patches inherently prefer external denoising. Such a preference is in essence the tradeoff between noise-fitting versus signal-fitting. Burger et al. [15] proposed a learning-based approach that automatically combines denoising results from an internal and an external method. The learned combining strategy outperforms both internal and external approaches across a wide range of images, being closer to theoretical bounds.

In super-resolution literature, while the most popular methods are based on either external or internal similarities, there have been previous efforts to utilize one to regularize the other. For example, the method proposed in [18] incorporated both a local autoregressive (AR) model and a nonlocal self-similarity regularization term, into the sparse representation framework. The local self-similarity is further refined by Yang et al. [19] using in-place self-similarity, meaning that patch matching can be restricted to its original location in the lower-scale image. In order to handle these in-place internal examples, the authors learned a robust first-order approximation of the nonlinear mapping function from a collection of external images. The algorithm can produce more natural structures and is thereby better at handling real applications.

CHAPTER 3

A JOINT SR MODEL

Let $\mathbf{X}$ denote the HR image to be estimated from the LR input $\mathbf{Y}$. $\mathbf{X}_{ij}$ and $\mathbf{Y}_{ij}$ stand for the (i,j)-th $(i,j = 1, 2...)$ patch from $\mathbf{X}$ and $\mathbf{Y}$, respectively. Considering almost all SR methods work on patches, we define two loss functions $\ell_{\mathcal{G}}(\cdot)$ and $\ell_{\mathcal{I}}(\cdot)$ in a patch-wise manner, which enforce the external and internal similarities for regularization, respectively. While one intuitive idea is to minimize a weighted combination of the two loss functions, a patch-wise (adaptive) weight $\omega(\cdot)$ is needed to balance them. We now write our proposed joint SR in the following form:

$$\min_{\mathbf{x}_{ij},\alpha_{ij},\mathbf{X}_{ij}^E} \quad \ell_{\mathcal{G}}(\mathbf{X}_{ij}, \mathbf{a}_{ij}) + \omega(\mathbf{a}_{ij}, \mathbf{X}_{ij}^E)\ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}_{ij}^E). \tag{3.1}$$

We will discuss the formulation of each component in (3.1) in the next sections.

We will only discuss one specific form of joint SR in this thesis. However, note that with different choices of $\ell_{\mathcal{G}}(\cdot)$, $\ell_{\mathcal{I}}(\cdot)$, and $\omega(\cdot)$, a variety of methods can be accommodated in the framework. For example, if we set $\ell_{\mathcal{G}}(\cdot)$ as the (adaptively reweighted) sparse coding term, while choosing $\ell_{\mathcal{I}}(\cdot)$ equivalent to the two local and non-local similarity based terms, then (3.1) becomes the model proposed in [18], with $\omega(\cdot)$ being some empirically chosen constants.

3.1 Sparse Coding for External Examples

The HR and LR patch spaces $\{\mathbf{X}_{ij}\}$ and $\{\mathbf{Y}_{ij}\}$ are assumed to be tied by some mapping function. With a well-trained coupled dictionary pair ($\mathbf{D_h}$, $\mathbf{D_l}$) (see [4] for details on training a coupled dictionary pair), the *coupled sparse coding*

[9] assumes that $(\mathbf{X}_{ij}, \mathbf{Y}_{ij})$ tends to admit a joint sparse representation $\mathbf{a}_{ij}$:

$$\min_{\mathbf{X}_{ij},\mathbf{a}_{ij}} \quad \lambda||\mathbf{a}_{ij}||_1 + ||\mathbf{D_l}\mathbf{a}_{ij} - \mathbf{Y}_{ij}||_F^2 + ||\mathbf{D_h}\mathbf{a}_{ij} - \mathbf{X}_{ij}||_F^2. \tag{3.2}$$

Since $\mathbf{X}$ is unknown, Yang et al. [9] suggest to first infer the sparse code $\mathbf{a}_{ij}^L$ of $\mathbf{Y}_{ij}$ with respect to $\mathbf{D_l}$, and then use it as an approximation of $\mathbf{a}_{ij}^H$ (the sparse code of $\mathbf{X}_{ij}$ with respect to $\mathbf{D_h}$), to recover $\mathbf{X}_{ij} \approx \mathbf{D_h}\mathbf{a}_{ij}^L$. We are targeted at the loss function itself, with the external similarity enforced by coupled dictionaries:

$$\ell_{\mathcal{G}}(\mathbf{X}_{ij}, \mathbf{a}_{ij}) = \lambda||\mathbf{a}_{ij}||_1 + ||\mathbf{D_l}\mathbf{a}_{ij} - \mathbf{Y}_{ij}||_F^2 + ||\mathbf{D_h}\mathbf{a}_{ij} - \mathbf{X}_{ij}||_F^2. \tag{3.3}$$

3.2 Epitome Matching for Internal Examples

3.2.1 The High Frequency Transfer Scheme

Freedman and Fattal [5] observed that small patches, especially singular features like edges and corners, tend to repeat almost identically across different image scales. Their "high frequency transfer" method searches the high-frequency component for a target HR patch, by NN patch matching across scales. Defining a linear interpolation operator $\mathcal{U}$ (with scaling factor s) and a downsampling operator $\mathcal{D}$ (with scaling factor $\frac{1}{s}$), for the input LR image $\mathbf{Y}$, we first obtain its initial upsampled image $\mathbf{X}'^E = \mathcal{U}(\mathbf{Y})$, and a smoothed input image $\mathbf{Y}' = \mathcal{D}(\mathcal{U}(\mathbf{Y}))$. Given the smoothed patch $\mathbf{X}_{ij}'^E$, the missing high-frequency band of each unknown patch $\mathbf{X}_{ij}^E$ is predicted by first solving a NN matching (3.4):

$$(m, n) = \arg\min_{(m,n)\in\mathcal{W}_{ij}} \|\mathbf{Y}'_{mn} - \mathbf{X}_{ij}'^E\|_F^2, \tag{3.4}$$

where $\mathcal{W}_{ij}$ is defined as a small searching window centered at $(\frac{i}{s}, \frac{j}{s})$ on image $\mathbf{Y}'$. We could also simply express it as $(m, n) = f_{NN}(\mathbf{X}_{ij}'^E, \mathbf{Y})$.

With the co-located patch $\mathbf{Y}_{mn}$ from $\mathbf{Y}$, the high-frequency band $\mathbf{Y}_{mn} - \mathbf{Y}'_{mn}$ is pasted onto $\mathbf{X}_{ij}'^E$, i.e., $\mathbf{X}_{ij}^E = \mathbf{X}_{ij}'^E + \mathbf{Y}_{mn} - \mathbf{Y}'_{mn}$.

3.2.2 The Epitomic Matching Algorithm

The matching of $\mathbf{X}'^{E}_{ij}$ over the smoothed input image $\mathbf{Y}'$ makes the core step of the high frequency transfer scheme. However, the performance of NN matching (3.4) is degraded with the presence of noise and outliers. Moreover, the NN matching in [5] is restricted to a local window for efficiency, which potentially accounts for some rigid artifacts.

Instead, we propose *epitomic matching* to replace NN matching in the above frequency transfer scheme. As a generative model, epitome [22] summarizes a large set of raw image patches into a condensed representation. We first learn an epitome $\mathbf{e}_{\mathbf{Y}'}$ from $\mathbf{Y}'$, and then match each $\mathbf{X}'^{E}_{ij}$ over $\mathbf{e}_{\mathbf{Y}'}$ rather than $\mathbf{Y}'$ directly. Assume $(m, n) = f_{ept}(\mathbf{X}'^{E}_{ij}, \mathbf{Y}, \mathbf{e}_{\mathbf{Y}'})$, where f_{ept} denotes the procedure of epitomic matching. The high frequency transfer scheme is then performed in the same way as [5]: $\mathbf{X}^{E}_{ij} = \mathbf{X}'^{E}_{ij} + \mathbf{Y}_{mn} - \mathbf{Y}'_{mn}$: the only difference here lies between f_{ept} and f_{NN}.

The algorithm procedure of epitomic matching goes as follows. We assume an epitome $\mathbf{e}$ of size $M_e \times N_e$, for an input image of size $M \times N$, where $M_e < M$ and $N_e < N$. Similarly to GMMs, $\mathbf{e}$ contains three parameters [20, 21, 22]: $\boldsymbol{\mu}$, the Gaussian mean of size $M_e \times N_e$; $\boldsymbol{\phi}$, the Gaussian variance of size $M_e \times N_e$; and $\boldsymbol{\pi}$, the mixture coefficients. Suppose there are Q densely sampled, overlapped patches from the input image, i.e. $\{\mathbf{Z}_k\}_{k=1}^{Q}$. Each $\mathbf{Z}_k$ contains pixels with image coordinates $\mathbf{S}_k$, and is associated with a hidden mapping $\mathcal{T}_k$ from $\mathbf{S}_k$ to the epitome coordinates. All the Q patches are generated independently from the epitome and the corresponding hidden mappings as below:

$$\prod_{k=1}^{Q} p(\{\mathbf{Z}_k\}_{k=1}^{Q} | \{\mathcal{T}_k\}_{k=1}^{Q}, \mathbf{e}) = \prod_{k=1}^{Q} p(\mathbf{Z}_k | \mathcal{T}_k, \mathbf{e}). \tag{3.5}$$

The probability $p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e})$ in (3.5) is computed by the Gaussian distribution where the Gaussian component is specified by the hidden mapping $\mathcal{T}_k$. The behavior of $\mathcal{T}_k$ is similar to that of the hidden variable in the traditional GMMs.

Figure 3.1 illustrates the role that the hidden mapping plays in the epitome as well as the graphical model illustration for epitome. With all the above notations, our goal is to find the epitome $\mathbf{e}$ that maximizes the log

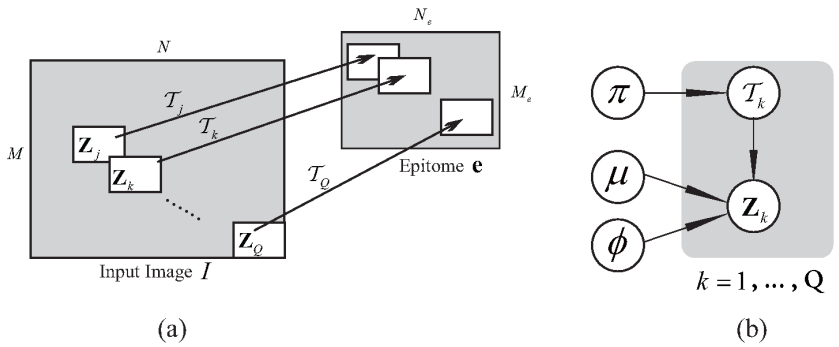


Figure 3.1: (a) The hidden mapping $\mathcal{T}_k$ maps the image patch $\mathbf{Z}_k$ to its corresponding patch of the same size in $\mathbf{e}$, and $\mathbf{Z}_k$ can be mapped to any possible epitome patch in accordance with $\mathcal{T}_k$. (b) The epitome graphical model.

likelihood function $\mathbf{e} = \arg\max_{\mathbf{e}} \log p\left(\{\mathbf{Z}_k\}_{k=1}^{Q}|\mathbf{e}\right)$, which can be solved by the Expectation-Maximization (EM) algorithm [20, 23].

With the epitome $\mathbf{e}_{\mathbf{Y}'}$ learned from the smoothed input image $\mathbf{Y}'$, the location of the matching patch in the epitome $\mathbf{e}_{\mathbf{Y}'}$ for each patch $\mathbf{X}'^{E}_{ij}$ is specified by the most probable hidden mapping for $\mathbf{X}'^{E}_{ij}$:

$$\mathcal{T}^*_{ij} = \arg\max_{\mathcal{T}_{ij}} p\left(\mathcal{T}_{ij}|\mathbf{X}'^{E}_{ij}, \mathbf{e}\right). \tag{3.6}$$

The patches in $\mathbf{Y}'$ with large posterior probabilities $p\left(\mathcal{T}^*_{ij}|\cdot, \mathbf{e}\right)$ are regarded as the candidate matches for the patch $\mathbf{X}'_{ij}$.

3.2.3 Epitome Matching for Internal SR

Note that each epitome patch summarizes a batch of similar raw patches in $\mathbf{Y}'$. For any patch $\mathbf{Y}'_{ij}$ that contains certain noise or outliers in $\mathbf{Y}'$, its posterior would be small, and it thus tends not be selected as candidate matches for $\mathbf{X}'^{E}_{ij}$, improving the robustness of matching. In addition, the epitome summarizes the patches of the entire $\mathbf{Y}'$, which refer to not only a local neighborhood but also non-local examples. In Chapter 4, we will show information about the performance comparison between internal SR using NN and epitomic matching. Finally, we define

$$\ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}^{E}_{ij}) = ||\mathbf{X}_{ij} - \mathbf{X}^{E}_{ij}||^2_F, \tag{3.7}$$

where $\mathbf{X}_{ij}^E$ is the internal SR result given by epitomic matching.

3.3 Learning the Adaptive Weights

In [17], Mosseri et al. showed that the internal versus external preference is tightly related to the Signal-to-Noise-Ratio (SNR) estimate of each patch. Inspired by that finding, we propose to define and estimate the patch-wise "SR noise" introduced by external and internal methods. The *external noise* is defined by the residual of sparse coding

$$N_g(\alpha_{ij}) = ||\mathbf{D_l}\mathbf{a}_{ij} - \mathbf{Y}_{ij}||_F^2. \tag{3.8}$$

Meanwhile, the *internal noise* finds its counterpart definition by the epitome matching error within f_{pet}:

$$N_i(\mathbf{X}_{ij}^E) = ||\mathbf{Y'}_{mn} - \mathbf{X}_{ij}^{'E}||_F^2, \tag{3.9}$$

where $\mathbf{Y}'_{mn}$ is the matching patch in $\mathbf{Y}'$ for $\mathbf{X}_{ij}^{'E}$.

Usually, the two noises are on the same magnitude level, which aligns with the fact that external- and internal-examples will have similar performances on the given patch (homogenous regions, etc.). However, there do exist patches where the two have a significant difference, which means the patch has a strong preference toward one of them. In such cases, we hope the "preferred" term can be sufficiently emphasized, and thus construct the following patch-wise adaptive weight (p is the controlling parameter):

$$\omega(\alpha_{ij}, \mathbf{X}_{ij}^E) = \exp(p \cdot [N_g(\mathbf{a}_{ij}) - N_i(\mathbf{X}_{ij}^E)]). \tag{3.10}$$

When the internal noise becomes larger, the weight decays quickly to ensure that external similarity dominates, and vice versa.

3.4 Algorithm

Directly solving (3.1) is very complex due to the its high nonlinearity and entanglement among all variables. Instead, we follow the coordinate descent

fashion [24] and solve the following three sub-problems iteratively.

3.4.1 $\mathbf{a}_{ij}$-subproblem

Fixing $\mathbf{X}_{ij}$ and $\mathbf{X}_{ij}^E$, we have the following minimization w.r.t α_{ij}

$$\begin{aligned}\min_{\mathbf{a}_{ij}} \quad & \lambda||\mathbf{a}_{ij}||_1 + ||\mathbf{D_l}\mathbf{a}_{ij} - \mathbf{Y}_{ij}||_F^2 + ||\mathbf{D_h}\mathbf{a}_{ij} - \mathbf{X}_{ij}||_F^2 \\ & +[\ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}_{ij}^E) \cdot \exp(-p \cdot N_i(\mathbf{X}_{ij}^E)] \cdot \exp(p \cdot N_g(\mathbf{a}_{ij})).\end{aligned} \tag{3.11}$$

The major bottleneck of exactly solving (3.11) lies in the last exponential term. We let $\mathbf{a}_{ij}^0$ denote the $\mathbf{a}_{ij}$ value solved in the last iteration. We then apply first-order Taylor expansion to the last term of the objective in (3.11), with regard to $N_g(\alpha_{ij})$ at $\alpha_{ij} = \alpha_{ij}^0$, and solve the approximated problem as follows:

$$\min_{\mathbf{a}_{ij}} \quad \lambda||\mathbf{a}_{ij}||_1 + (1+C)||\mathbf{D_l}\mathbf{a}_{ij} - \mathbf{Y}_{ij}||_F^2 + ||\mathbf{D_h}\mathbf{a}_{ij} - \mathbf{X}_{ij}||_F^2, \tag{3.12}$$

where C is the constant coefficient:

$$\begin{aligned}C &= [\ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}_{ij}^E) \cdot \exp(-p \cdot N_i(\mathbf{X}_{ij}^E)] \cdot [p \cdot \exp(p \cdot N_g(\mathbf{a}_{ij}^0)] \\ &= p\ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}_{ij}^E) \cdot \omega(\alpha^{\mathbf{0}}{}_{ij}, \mathbf{X}_{ij}^E).\end{aligned} \tag{3.13}$$

Equation (3.12) can be conveniently solved by the feature sign algorithm [8]. Note that (3.12) is a valid approximation of (3.11) since $\mathbf{a}_{ij}$ and $\mathbf{a}_{ij}^0$ become quite close after a few iterations, so that the higher-order Taylor expansions can be reasonably ignored.

Another noticeable fact is that since $C > 0$, the second term is always emphasized more than the third term, which makes sense as $\mathbf{Y}_{ij}$ is the "accurate" LR image, while $\mathbf{X}_{ij}$ is just an estimate of the HR image and is thus less weighted. Further considering the formulation (3.13), C grows up as $\omega(\alpha^{\mathbf{0}}{}_{ij}, \mathbf{X}_{ij}^E)$ turns larger. That implies when external SR becomes the major source of "SR noise" on one patch, then correspondingly (3.12) will automatically rely less on the last $\mathbf{X}_{ij}$.

3.4.2 $\mathbf{X}_{ij}^E$-subproblem

Fixing $\mathbf{a}_{ij}$ and $\mathbf{X}_{ij}$, the $\mathbf{X}_{ij}^E$-subproblem becomes

$$\min_{\mathbf{X}_{ij}^E} \quad \exp(-p \cdot ||\mathbf{Y}'_{mn} - \mathbf{X}_{ij}'^E||_F^)\ell_{\mathcal{I}}(\mathbf{X}_{ij}, \mathbf{X}_{ij}^E). \tag{3.14}$$

Note that (3.14) is in essence a reweighed version of epitomic matching described in Section 3.2, and could be solved by similar algorithms.

3.4.3 $\mathbf{X}_{ij}$-subproblem

With both $\mathbf{a}_{ij}$ and $\mathbf{X}_{ij}^E$ fixed, the solution of $\mathbf{X}_{ij}$ simply follows a weight least square (WLS) problem:

$$\min_{\mathbf{X}_{ij}} \quad ||\mathbf{D_h}\mathbf{a}_{ij} - \mathbf{X}_{ij}||_F^2 + \omega(\mathbf{a}_{ij}, \mathbf{X}_{ij}^E)||\mathbf{X} - \mathbf{X}_{ij}^E||_F^2, \tag{3.15}$$

with an explicit solution:

$$\mathbf{X}_{ij} = \frac{\mathbf{D_h}\mathbf{a}_{ij} + \omega(\alpha_{ij}, \mathbf{X}_{ij}^E) \cdot \mathbf{X}_{ij}^E}{1 + \omega(\mathbf{a}_{ij}, \mathbf{X}_{ij}^E)}. \tag{3.16}$$

CHAPTER 4

EXPERIMENTS

4.1 Implementation Details

We itemize the parameter and implementation settings for the following group of experiments:

- We use 5×5 patches with one pixel overlapping for all experiments except those on SHD images in Section 4.4, where the patch size is 25×25 with five pixels overlapping.
- In (3.3), we adopt the $\mathbf{D_l}$ and $\mathbf{D_h}$ trained in the same way as in [4], due to the similar roles played by the dictionaries in their formulation and our ℓ_G function. However, we are aware that such D_l and D_h are not optimized for the proposed method, and will integrate a specifically designed dictionary learning part in future work. λ is empirically set as 1.
- In (3.7), the size of the epitome is $\frac{1}{4}$ of the image size.
- In (3.13), we set $p = 1$ for all experiments. We also observed in experiments that a larger p will usually lead to a faster decrease in objective value, but the SR result quality may degrade a bit.
- We initialize $\mathbf{a}_{ij}$ by solving coupled sparse coding in [4]. $\mathbf{X}_{ij}$ is initialized by bicubic interpolation.
- We set the maximum iteration number to be 10 for the coordinate descent algorithm for a trade-off between accuracy and efficiency. For SHD cases, the maximum iteration number is adjusted to be 5.

- For color images, we apply SR algorithms to the illuminance channel only, as humans are more sensitive to illuminance changes. We then interpolate the color layers (Cb, Cr) using plain bi-cubic interpolation.

4.2 Effect of Adaptive Weight

To demonstrate how the proposed joint SR will benefit from the learned adaptive weight (3.13), we list comparisons between (3.1) and its fixed weight counterpart, i.e. set the weight ω as constant for all patches. Figure 4.1 shows that the joint SR with an adaptive weight gains a consistent PSNR advantage over the SR with a large range of fixed weights. Further, we visualize the patch-wise weight map of joint SR on the *Kid* image, at iterations 1, 3, 7, 5, and 10, as heat maps in Fig. 4.2. The weights start with a relatively scattered distribution at the very beginning. Yet during iterations, most homogenous regions, as well as the eyes and nose, are associated with rather small weights, showing that the external similarity is able to gain more advantages there. The larger weight values become sparse and significantly focused on the repetitive textures, e.g. the sweater, where the self-similarity will be imposed more heavily. Note the observation is in accordance with our intuition in Chapter 2. We also note that the weight map changes little from iteration 7 to iteration 10, which implies that the joint SR algorithm reaches a stable solution after a few iterations.

4.3 Comparison with State-of-the-Art Results

We compare the proposed method with the following selection of competitive methods:

- *Bi-Cubic Interpolation (BCI for short and similarly hereinafter)*, as a comparison baseline.
- *Coupled Sparse Coding (CSC)* [4], as the classical external-example-based SR.
- *Local Self-Example-based SR (LSE)* [5], as the classical internal-example-based SR.

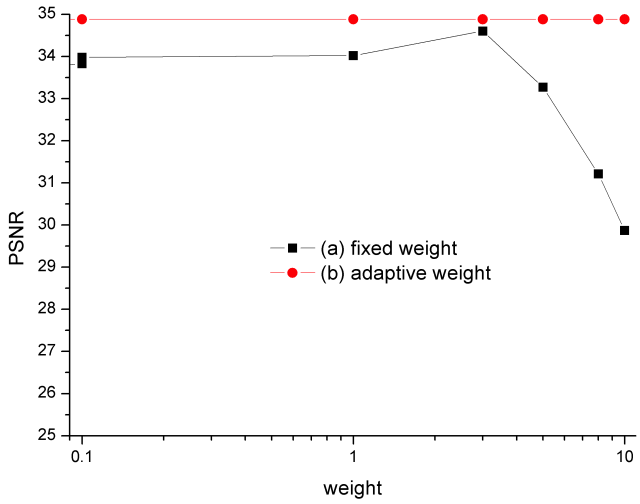


Figure 4.1: The PSNR curve of the *Kid* image: (a) with different fixed weights; (b) with an adaptive weight.

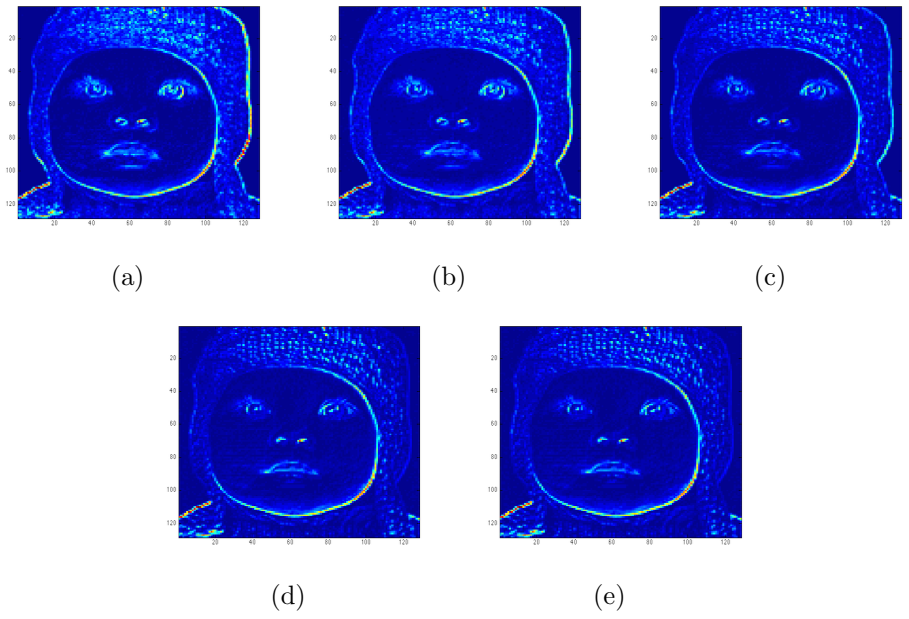
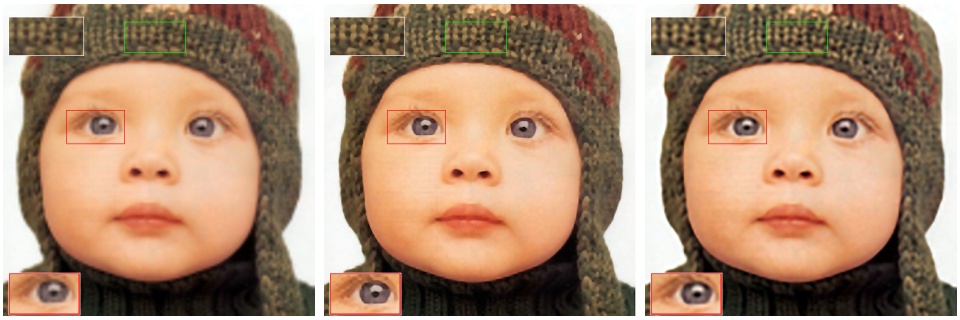


Figure 4.2: The weight map produced by joint SR on the *Kid* image at (a) iteration 1; (b) iteration 3; (c) iteration 5; (d) iteration 7; and (e) iteration 10.

- *Epitome-based SR (EPI)*, which follows the same high frequency transfer scheme of LSE [5], except for replacing the NN matching with our proposed epitomic matching. We include it in the comparison to show the merits of epitome matching over NN matching.
- *SR based on In-place Example Regression (IER)* [19], as the previous SR utilizing both external and internal information.
- *The proposed Joint SR (JSR).*



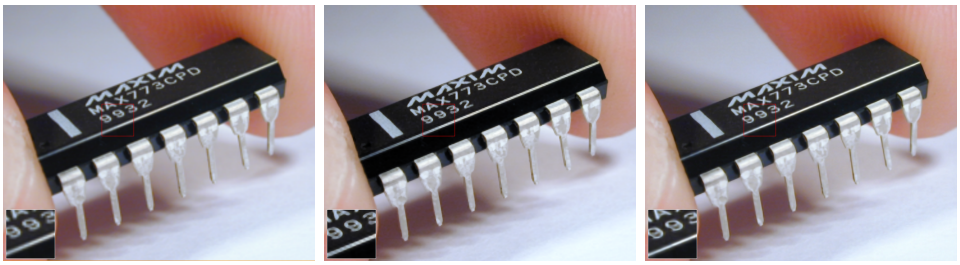
(a) BCI (c) LSE (e) IER

(b) CSC (d) EPI (f) JSR

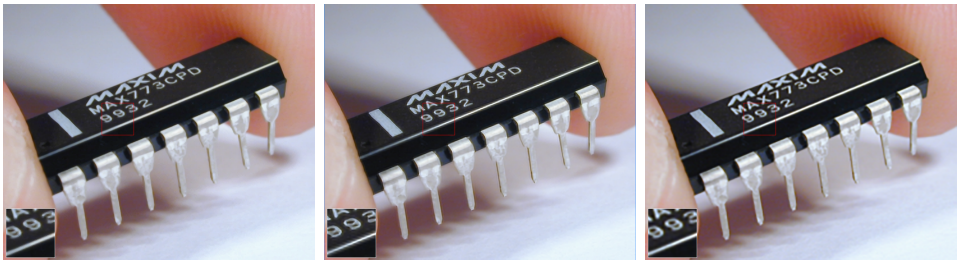
Figure 4.3: 4× SR result of the *Kid* image.

In this section, we magnify all the input LR images by a factor of 4. We generate SR results for CSC and IER using their original codes [4], [19], and implement LSE [5] to the best of our ability. As the visual quality is the most important criterion for evaluating SR, we list the visual comparison results for all the three test images (results are best viewed on a high-resolution display), from Fig. 4.3 to Fig. 4.5. Detailed comparisons can be found at the enlarged local regions.

Figure 4.3 shows the SR results for the *Kid* image. Although outperforming the naive BCI noticeably, the external-example-based CSC tends to



(a) BCI (c) LSE (e) IER



(b) CSC (d) EPI (f) JSR

Figure 4.4: 4× SR result of the *Chip* image.

(a) BCI (c) LSE (e) IER

(b) CSC (d) EPI (f) JSR

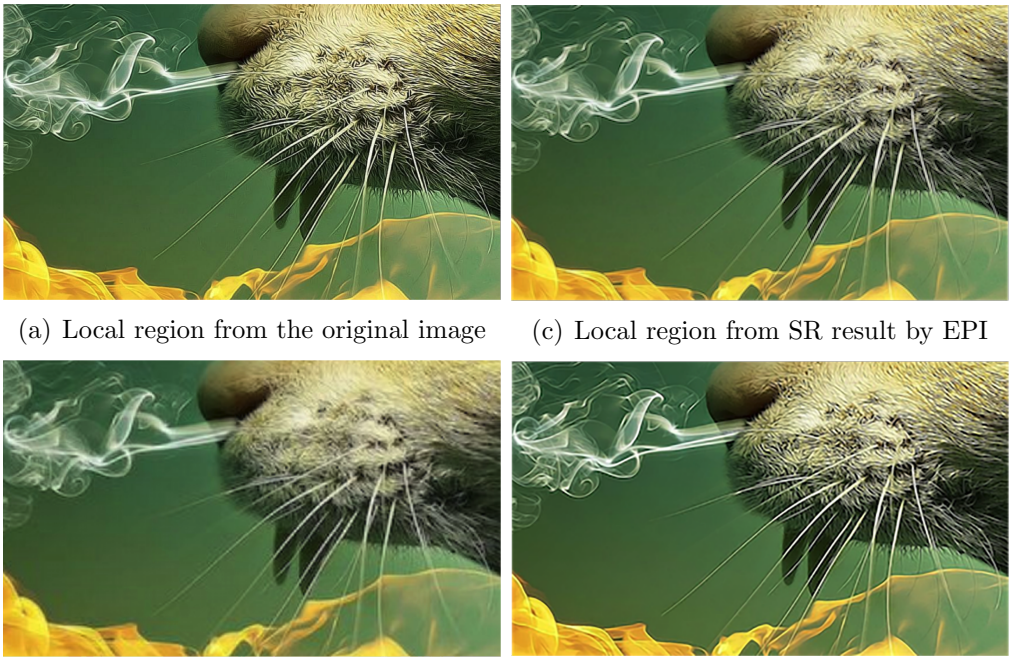
Figure 4.5: 4× SR result of the *Statue* image.

oversmooth many fine details along the textures in the hat and sweater, the eyelash, and so on. In contrast, LSE brings out an overly sharp SR result. Obvious blockiness is observable in its enlarged hat regions, inconsistency in eyelids, as well as ringings and jaggies along the edges of faces and eyeballs. Compared the LSE, EPI produces a more visually pleasing result with a part of the artifacts removed. But without any external reference information available, EPI is still incapable of inferring more high-frequency details from the input LR image solely, especially under a large amplifying factor. IER, being a joint SR method in essence, greatly improves in generating both an artifact-free and detail-preserving SR result. However, the textures are a little blurry with occasional small artifacts. One may even argue that its eye and hat parts are not as natural-looking as in CSC. Finally, the proposed JSR combines the advantages of both external and internal examples, and leads to the most satisfactory result. The algorithm adaptively emphasizes external examples for homogenous and regular regions (face, eyeballs, etc.), avoiding rigid artifacts caused by the bad matches of internal examples. For regions containing singular features (eyelash, etc.) that external examples can only lead to over smoothness, JSR automatically shows a more similar performance to LSE/EPI and gives out sharp, clear details resulting from internal examples. As a result, the textures on the child's hat and sweater are nicely preserved, while the fine structures of the eye parts are better reconstructed than others.

In Figure 4.4, the *Chip* image is very challenging for SR due to its abundance of edges and textures. The CSC result is generally a bit blurry, especially in the characters on the chip surface. Both LSE and EPI perform well at enhancing edges, but create jaggy artifacts along the long edge of the chip surface as well as small structure distortions. IER is superior in removing artifacts but it is also not sufficiently sharp. The JSR result presents the best reconstruction of the characters without any noticeable artifacts. Similar comparisons could be found in Fig. 4.5 of the *Statue* image, where the differences in SR performance can be easily identified in the enlarged textured areas.

4.4 SR Beyond Standard Definition: From HD Image to UHD Image

In almost all SR literature, experiments are conducted on Standard-Definition (SD) images (most often 720 × 480 or 720 × 576 pixels) or even smaller. Today, the TV industry supports two popular High-Definition (HD) formats; 720p (1280 × 720 pixels) and 1080p (1920 × 1080 pixels, Full HD). All HDTVs come with HD resolution. Moreover, the TV industry is already pushing for the Ultra High-Definition (UHD) standard, which covers both 4K/2160p (3840 × 2160 pixels) and 8K/4320p (7680 × 4320 pixels). UHD TVs are hitting the consumer markets right now with true 3840 × 2160 resolution. It is thus quite interesting to explore whether all those SR algorithms tested on SD images can also be applied or adjusted for HD or UHD cases. In this section, we upscale HD images of 1280 × 720 pixels to SHD results of 3840 × 2160 pixels, using several previous methods and our JSR algorithm.



(a) Local region from the original image (c) Local region from SR result by EPI

(b) Local region from SR result by CSC (d) Local region from SR result by JSR

Figure 4.6: 3× SR results of the *Leopard* image (local region displayed).

Since most HD and UHD images typically contain much more diverse textures and richer fine structures than SD images, we enlarge the patch size from 5 × 5 to 25 × 25 (the dictionary pair is therefore re-trained as well),

meanwhile increasing the overlapping from one pixel to five pixels for enough spatial consistency. Hereby JSR is compared with its two "component" algorithms, i.e., CSC and EPI. We choose several challenging SHD images (3840 $\times$ 2160 pixels) with very cluttered texture regions, downsampling them to HD size (1280 $\times$ 720 pixel) on which we apply the SR algorithm with a factor of 3. SR results on two HD images, *Leopard* and *Grass*, are displayed in Fig. 4.6 and Fig. 4.7 as zoomed local regions, while original results are included in supplementary materials (better viewed on UHD display). In both cases, our results are consistently sharper and clearer.

(a) Local region from the original image

(c) Local region from SR result by EPI

(b) Local region from SR result by CSC

(d) Local region from SR result by JSR

Figure 4.7: 3$\times$ SR results of the *Grass* image (local region displayed).

4.5 Subjective Evaluation

We conduct an online *subject evaluation*[1] on the quality of SR results produced by different methods. The methods under comparison include BIC, CSC, LSE, IER, EPI, and JSR. Ground truth HR images are also included when they are available as references. Each participant is shown a set of HR image pairs obtained using two different methods for the same LR image. For each pair, the participant needs to decide which one is better than the other in terms of perceptual quality. The image pairs are drawn from all the competitive methods randomly, and the images winning the pairwise comparison will be compared again in the next round, until the best one is selected.

We have a total of 101 participants giving 1,047 pairwise comparisons, over six images with different scaling factors (*Kid*×4, *Chip*×4, *Statue*×4, *Lion*×3, *Temple*×3 and *Train*×3). Not every participant completed all the comparisons but their partial responses are still useful. All the evaluation results can be summarized in a 7×7 winning matrix $\mathbf{W}$ for seven methods (including ground truth), based on which we fit a Bradley-Terry [25] model to estimate the subjective score for each method so that they can be ranked. In the Bradley-Terry model, the probability that an object X is favored over Y is assumed to be

$$p(X \succ Y) = \frac{e^{s_X}}{e^{s_X} + e^{s_Y}} = \frac{1}{1 + e^{s_Y - s_X}}, \tag{4.1}$$

where s_X and s_Y are the subjective scores for X and Y. The scores $\mathbf{s}$ for all the objects can be jointly estimated by maximizing the log likelihood of the pairwise comparison observations:

$$\max_{\mathbf{s}} \sum_{i,j} w_{ij} \log \left(\frac{1}{1 + e^{s_j - s_i}} \right), \tag{4.2}$$

where w_{ij} is the (i, j)-th element in the winning matrix $\mathbf{W}$, representing the number of times when method i is favored over method j. We use the Newton-Raphson method to solve (4.2) and set the score for ground truth as 1 to avoid the scale issue.

Figure 4.8 shows the estimated scores for the six SR methods in our eval-

[1] `http://www.ifp.illinois.edu/~wang308/survey`

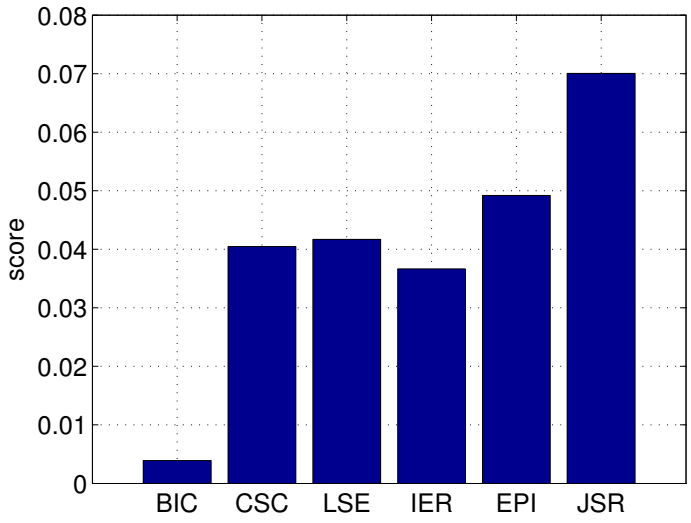


Figure 4.8: Subjective SR quality score for different methods. The ground truth has score 1.

uation. As expected, all SR methods receive much lower scores compared to ground truth, showing the huge challenge of the SR problem itself. Also, the bicubic interpolation is significantly worse than others. The proposed JSR method then outperforms all other state-of-the-art methods by a large margin, which verifies that JSR can produce more visually favorable HR images.

CHAPTER 5

CONCLUSION

This thesis contributes to a joint single image SR model, by learning from both external and internal examples. We define the two loss functions by sparse coding and epitomic matching, respectively, and construct the adaptive weight. Experimental results demonstrate that the joint SR outperforms existing state-of-the-art methods for various test images of different definitions and scaling factors, and is also significantly more favored by user perception. Our future work will integrate dictionary learning into the proposed scheme, as well as reducing the complexity of the algorithm.

REFERENCES

[1] S. C. Park, M. K. Park and K. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21-36, 2003.

[2] R. Fattal, "Image upsampling via imposed edge statistics," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 95-102, 2007.

[3] Z. Lin and H. Y. Shum, "Fundamental limits of reconstruction-based superresolution algorithms under local translation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 83-97, 2004.

[4] J. Yang, Z. Wang, Z. Lin, S. Chen and T. S. Huang, "Coupled dictionary training for image super-resolution," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3467-3478, 2012.

[5] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Transactions on Graphics*, vol. 28, no. 3, 2010.

[6] W. T. Freeman, T. R. Jones and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56-65, 2002.

[7] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1127-1133, 2010.

[8] H. Lee, A. Battle, R. Raina and A. Y. Ng, "Efficient sparse coding algorithms," in *Proceedings of Neural Information Processing Systems (NIPS)*, pp. 801-808, 2007.

[9] J. Yang, J. Wright, T. S. Huang and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861-2873, 2010.

[10] C. Yang, J. Huang and M. Yang, "Exploiting self-similarities for single frame super-resolution," in *Proceedings of Asian Conference on Computer Vision (ACCV)*, pp. 1807-1818, 2010.

[11] W. Dong, L. Zhang, G. Shi and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1620-1630, 2012.

[12] D. Glasner, S. Bagon and M. Irani, "Super-resolution from a single image," in *Proceeding of IEEE International Conference on Computer Vision (ICCV)*, pp. 349-356, 2009.

[13] J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman, "Non-local sparse models for image restoration," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 2272-2279, 2009.

[14] P. Chatterjee and P. Milanfar, "Is denoising dead?" *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 895-911, 2010.

[15] H. Burger, C. Schuler and S. Harmeling,"Learning how to combine internal and external denoising methods," *Lecture Notes in Computer Science: Pattern Recognition*, vol. 8142, pp. 121-130, 2013.

[16] M. Zontak, "Internal statistics of a single natural image," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 977-984, 2011.

[17] I. Mosseri, M. Zontak and M. Irani, "Combining the power of internal and external denoising," *IEEE International Conference on Computational Photography (ICCP)*, pp. 1-9, 2013.

[18] W. Dong, D. Zhang, G. Shi and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1838-1857, 2011.

[19] J. Yang, Z. Lin and S. Cohen, "Fast image super-resolution based on in-place example regression," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1059-1066, 2013.

[20] N. Jojic, B. J. Frey and A. Kannan, "Epitomic analysis of appearance and shape," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 34-43, 2003.

[21] K. Ni, A. Kannan, A. Criminisi and J. Winn, "Epitomic location recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2158-2167, 2009.

[22] X. Chu, S. Yan, L. Li, K. Chan and T. S. Huang, "Spatialized epitome and its applications," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 311-318, 2010.

[23] Y. Yang, X. Chu, T.-T. Ng, A. Y-S. Chia, J. Yang, H. Jin and T. S. Huang, "Epitomic image colorization," in *Proceedings of IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[24] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Nashua, NH: Athena Scientific, 1999.

[25] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, pp. 324-345, 1952.